SOLUTION BRIEF

SUPERMICRO AND NVIDIA BRING HIGH END AI AND HPC DEVELOPMENT SYSTEMS TO OFFICE ENVIRONMENTS

Advanced System Reduces Power Consumption and Noise Levels While Delivering Massive AI and HPC Compute Performance

SYS-751GE-TNRT-NV1

TABLE OF CONTENTS

Executive Summary

AI is quickly becoming a mainstream technology used in a wide variety of industries. While the current view is that servers that excel at AI training must reside in a controlled environment in a data center, this new and innovative Supermicro Liquid Cooled AI development system with powerful CPUs and GPUs allows a much larger set of data scientists, engineers, and business analysts to make better decisions while reducing OPEX costs. Supermicro is advancing the state of AI by offering an AI server with state-of-the-art CPUs and GPUs with liquid cooling innovations that reduce power consumption and decibel levels. In addition, with the additional purchase of the optional NVIDIA AI Enterprise software and services, the SYS-751GE-TRT-NV1 is a complete solution aimed at the AI development professional. With the purchase of an optional subscription to NVIDIA AI Enterprise software, this unique system is ready to go, allowing developers and users to become productive in less time than ever before.

AI and HPC Use Cases

AI is being developed and used in a wide range of industries and workloads, including (but not limited to):

Figure 1 - AI and HPC Use Cases

AI Development and Execution Where Developers and Users Live

AI is becoming widespread, and developers require local systems with complete software and hardware control to create or execute new applications. The SYS-751GE-TNRT-NV1is a solution containing the necessary hardware for AI development and subsequent execution of applications with the optional purchase of NVAIE software.

Locality: Many of these domains and use cases require very low latency for interactive use. With the ability to quickly move this system with the combination of fast CPUs with multiple CPUs and a graphics accelerator, latencies are kept to a minimum. They do not involve relatively slow networking from an information worker's office. Although the hardware could be installed in a data center, the latencies to the users' screen would involve the transmission of graphics data over the network (VDI), which would not allow the combination of AI algorithms and display applications to work together. Locating the development system where the human is situated increases productivity for AI-based applications that require fast response times with minimal latencies.

Noise: The Liquid Cooled AI development platform reduces noise significantly. Compared to a data center, work areas are expected to have reduced noise, with the SYS-751GE-TNRT-NV1 AI development platform putting out about 30 dB when idle, 40dB when running at a 50% CPU load, and 50dB at a 100% load. The key is to utilize a self-contained liquid cooling system, where the CPUs and GPUs are all liquid cooled.

Mobility: Unlike a server in a data center, the SYS-751GE-TNRT-NV1 can move from one office to another. When a department's need changes, a powerful AI computing system can be easily redeployed with a simple connection to the network. If the AI development server needs to be placed in a centralized computing facility, the system can be easily mounted in a data center rack. This system is designed to be located in an office, cubicle, or at home due to noise levels of about 30dB and standard office electrical power requirements.

NVIDIA AI Enterprise Development Platform Summary

NVIDIA AI Enterprise is a complete set of AI software that allows developers and organizations to become more productive faster. In addition, this included software enables organizations to increase operational efficiency. With a full stack of AI software, including AI solution workflows, frameworks, and pre-trained models, the NVIDIA AI Enterprise software suite is a critical software component for AI developers and users. NVIDIA AI Enterprise is available on NVIDIA NGC, and this system, has the option to purchase the NVIDIA AI Enterprise subscription, at an additional cost, that provides access to an extensive library of full-stack software, including AI workflows, frameworks, and over 50+ NVIDIA pre-trained models, so organizations can develop once and run anywhere.

- Leverage fully integrated, optimized, certified, and supported software from NVIDIA for AI workloads.
- Run NVIDIA AI frameworks and tools optimized for GPU acceleration, reducing deployment time and ensuring reliable performance.
- Deploy anywhere – including on popular data center platforms from VMware and Red Hat, mainstream NVIDIA-Certified Systems configured with or without GPUs, and GPU-accelerated instances in the public cloud.
- Leverage the jointly certified NVIDIA and Red Hat solutions to deploy and manage AI workloads in containers or VMs with optimized software.
- Scale out to multiple nodes, enabling even the largest deep-learning training models to run on the VMware vSphere. Previously, scaling with bare metal performance in a fully virtualized environment was limited to a single node, limiting the complexity and size of AI workloads that could be supported.
- Run AI workloads at near bare-metal performance with new optimizations for GPU acceleration on vSphere, including support for the latest Ampere architecture, including the NVIDIA A100. Additionally, technologies like GPUDirect Communications can now be supported on vSphere. This ability provides communication between GPU memory and storage across a cluster for improved performance.

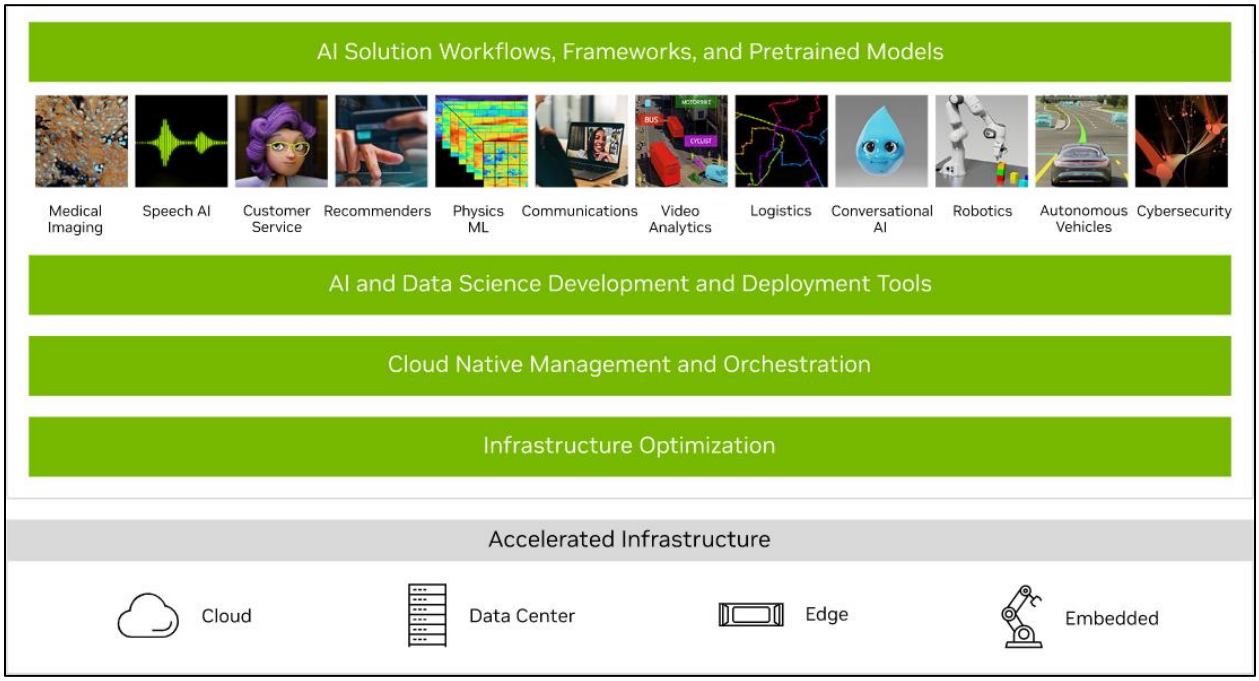

Figure 2 - NVIDIA AI Enterprise Software Stack (Image Courtesy of NVIDIA)

AI Development System Hardware Components

The Supermicro SYS-751GE-TNRT-NV1 is a complete solution that contains a number of powerful technologies that have been selected for the ultimate user experience:

- CPUs: Dual 4th Gen Intel® Xeon® Gold 6444Y, 16C/32T, running at a base clock rate of 3.6 GHz, with an all-turbo boost clock of 4.0 GHz. Intel Xeon CPU Max Series is also available.
- GPUs: 4x NVIDIA A100-LC 80GB PCIe GPUs (Liquid Cooled)
- GPU Interconnect: 2x NVIDIA NVLINK Bridge
- Memory: 512GB DDR5-4800MHz memory
- Storage: 6x Micron 1.9TB NVMe SSDs,
    - Two configured for RAID1 for the Operating System
    - Four for data storage.
- Graphics: NVIDIA Quadro RTX A4000
- NVIDIA Mellanox Conenct-X6 25Gb SFP28

Liquid Cooled AI Development System Details

A tightly packed system with two CPUs and four high-end GPUs creates a cooling problem that would be difficult to solve with typical server-level fans. The Supermicro SYS-751GE-TNRT-NV1 system is a self-contained liquid cooled system that rarely needs maintenance.

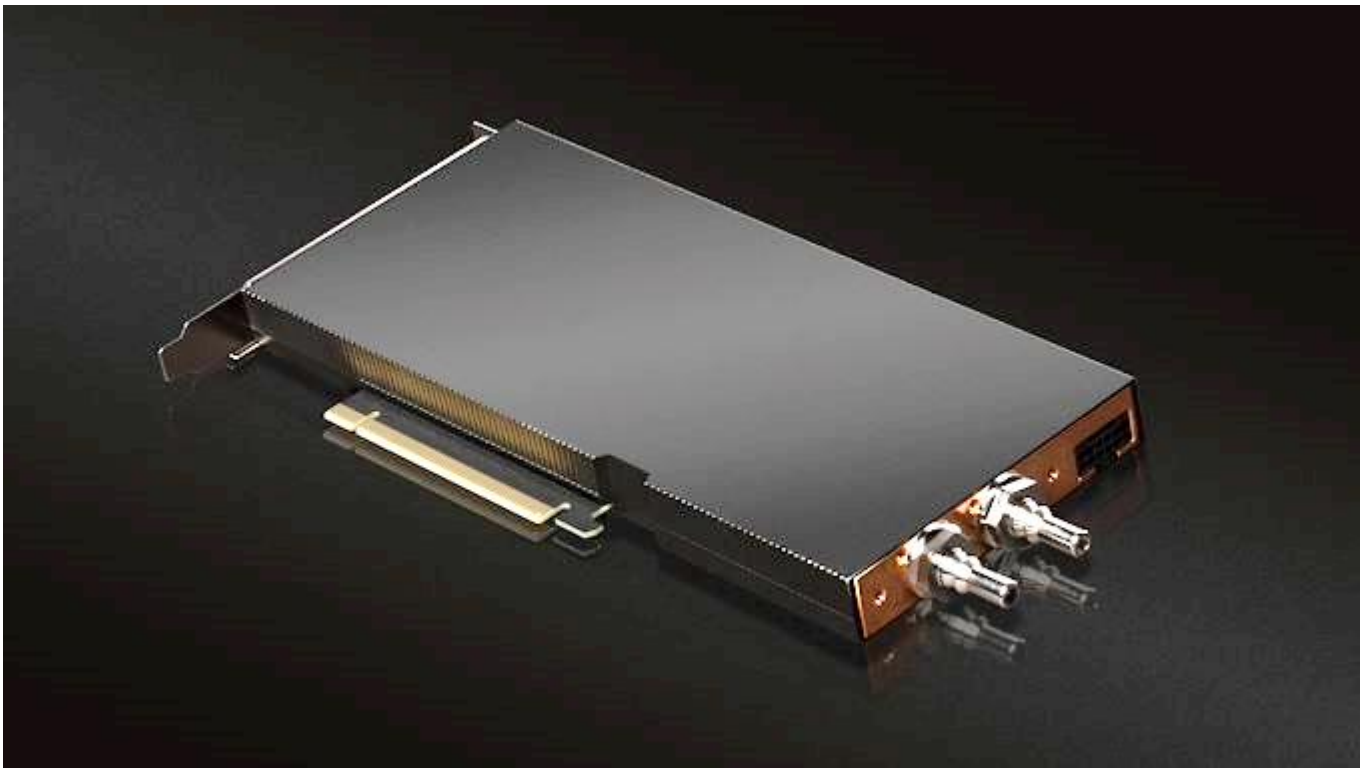

Figure 3 - NVIDIA Liquid Cooled A100

The figure below highlights the CPUs, Memory, PCIe Slots, Radiator, and Drive cage.

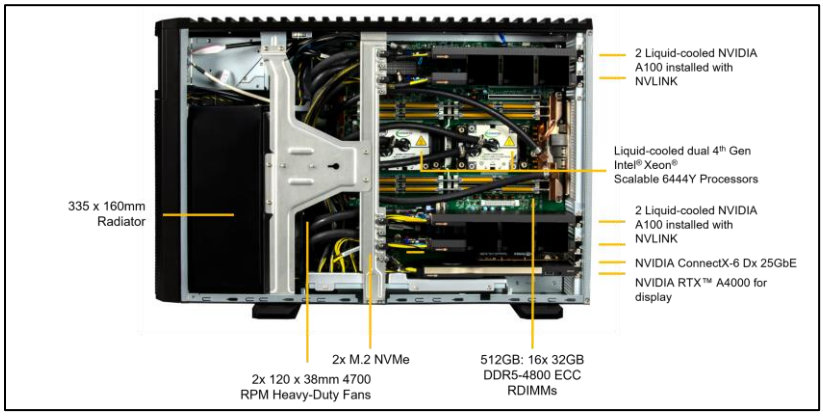

Figure 4 - Side view of Liquid Cooled System

Below is the front view of the system.

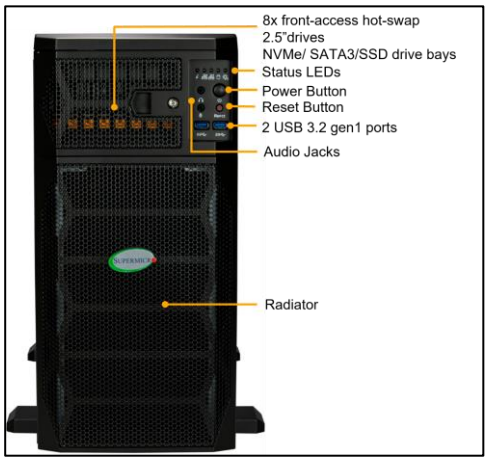

Figure 5 - Front View of Liquid Cooled System

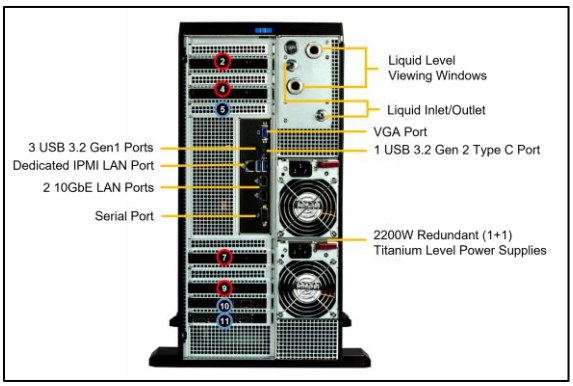

Figure 6 - Rear View of Liquid Cooled System

The open, side view of the SYS-751GE-TNRT-NV1

Figure 7 - Open Side View of Liquid Cooled System

AI Development Systems Liquid Cooling Components

A liquid cooled deskside system will typically not have access to the building infrastructure for the required process of cooling the liquid. Therefore, a self-contained system must be implemented within the server itself. The components that are needed include:

Cold Plates – these devices transfer the heat from the top of the hot CPU to the liquid. The liquid exits the cold plate warmer than when it entered. Cold plates can be set up to work in parallel.

Pumps – the liquid must be moved around in the system, from the reservoir to the cold plates to the radiator and back to the reservoir.

Hoses – Proper hose length and diameter are critical to the efficiency of a liquid-cooled development system.

Radiator – the heat from the CPUs and GPUs must be removed from the liquid. The simplest way to do this is using a simple radiator, where the warm liquid enters at the top of the radiator, and room temperature air is circulated over the liquid, reducing the temperature of the liquid as it exits the device.

Reservoir – The liquid cooling reservoir in the system stores extra coolant should it be needed.

Coolant – The Supermicro Liquid Cooling AI Development system contains a coolant developed to transport more heat from the CPUs and GPUs than other types of liquid.

Liquid Flow

The flow of the coolant in the Supermicro AI Development system is critical to the proper functioning of the overall system and takes the following path:

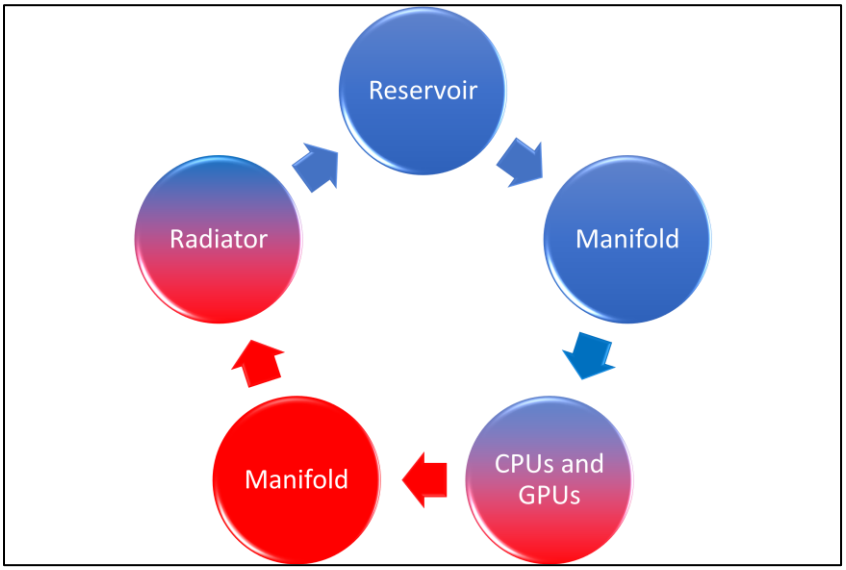

Figure 8 - Liquid Flow Path Through the Liquid Cooled System

Supermicro Liquid Cooling AI Development System Fluid Flow

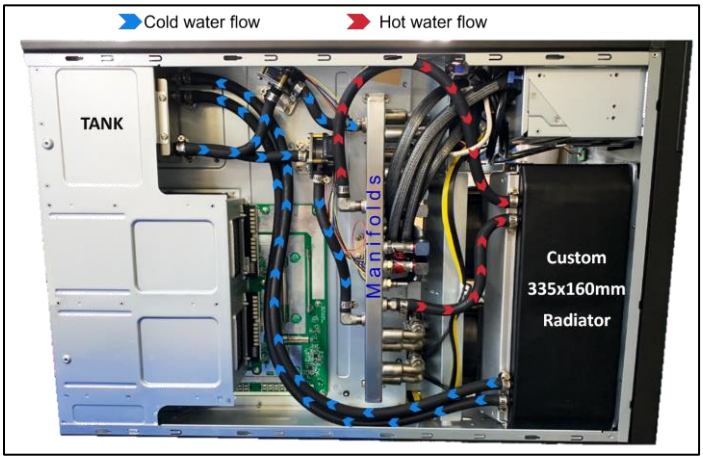

Figure 9 - Internal Liquid Flow Path

Supermicro AI Product Lines

Supermicro designs and delivers a wide range of servers with GPUs for AI and HPC acceleration. The Supermicro AI Development System is ideal for developing applications at a 4 GPU scale, and Supermicro designs and manufactures a wide range of GPU servers for deployment. When an application is ready for deployment, rack scale solutions are required. With servers containing up to 8 GPUs, and when networked together, the performance may be achieved multiple times in the deployment phase.

The Supermicro AI Product Line consists of servers in 1U, 2U, 4U, 6U, and 8U form factors. Using a range of GPUs from leading suppliers and with either PCIe, SXM, or OAM fabrics, up to 10 GPUs can be fitted into a single server.

Summary

The Supermicro SYS-751GE-TNRT-NV1 is an AI development and execution powerhouse. This system, located in an office or home environment, gives users over 2 Petaflops of AI performance. Built with the latest 4th Gen Intel Xeon Scalable processors and high-performance NVIDIA GPUs, this super quiet AI Development system provides data scientists, analytics engineers, and others the performance needed to use AI for any workload. In addition, with liquid cooling, the system allows for high performance CPUs and GPUs to be utilized without the typically associated noise levels.

For More Information:

Supermicro AI Development System – https://www.supermicro.com/en/products/system/gpu/tower/sys-751ge-tnrt-nv1

Supermicro GPU Servers - https://www.supermicro.com/en/products/gpu

NVIDIA AI Enterprise - https://www.nvidia.com/en-us/data-center/products/ai-enterprise/

Linus Tech Tips Video: https://www.youtube.com/watch?v=aSxomAgD8s4